Contrastive Learning and Co-Attention for Hierarchical Multi-Label Patent Classification

Xin Hu[1], Huaxiong Zhang[1,*], Zhijian Fang[1], Lei Zhou[2]
[1]School of Computer Science and Technology (School of Artificial Intelligence),
Zhejiang Sci-Tech University, Zhejiang, China.
[2]Zhejiang Talent Development Group.
[*]Corresponding author. Email:zhxhz@zstu.edu.cn

***Abstract*—Patent classification aims to automatically assign one or more IPC codes to patent documents. To address representation challenges caused by the highly technical nature of patent language and the prevalence of self-defined terminology, as well as the insufficient modeling of label semantics–text associations, we propose a hierarchical multi-label classification framework. Specifically, we weight contrastive constraints by hierarchical label-path similarity to avoid treating semantically close samples as negatives; we further design a co-attention mechanism to fuse label semantics with patent text and strengthen fine-grained matching signals, thereby improving the model's ability to characterize category decision boundaries. Experiments on USPTO-200K show statistically significant gains over PatCLS with up to 2.63% relative improvement in NDCG@3.**

***Index Terms*—Patent classification, Hierarchical multi-label, Contrastive learning, Co-attention mechanism.**

I. Introduction

The World Intellectual Property Organization (WIPO) reported in *World Intellectual Property Indicators 2025* that global patent filings exceeded 3.7 million in 2024, a 4.9% year-on-year increase. After receiving applications, intellectual property offices must assign one or more International Patent Classification (IPC) codes. This procedure is traditionally conducted by expert examiners, but manual classification is increasingly expensive and inefficient at scale [1]. Moreover, patent classification is challenging due to its multi-class and multi-label nature, and patent documents are filled with dense technical and legal terminology, imposing stringent requirements on examiners' domain expertise [2]. These factors motivate efficient and accurate hierarchical multi-label automated patent classification.

Recent studies on patent classification broadly follow three directions: (i) learning representations from text using conventional embeddings and neural encoders [3]–[5], (ii) fine-tuning pretrained language models for patent classification [6]–[8], and (iii) improving robustness via ensemble learning or multi-source information fusion [9]–[11]. Despite notable progress, most methods still rely predominantly on text representations and face two persistent challenges. First, patent language is highly technical and often introduces coined, self-defined terms for novel concepts, which are difficult for models pretrained on general-domain corpora to represent accurately. Second, many approaches insufficiently model fine-grained alignment between patent text and hierarchical labels, overemphasizing coarse document-level patterns while underutilizing subtle semantic cues at the term and sentence levels.

Meanwhile, contrastive learning has been widely applied to multi-label text classification, making it indispensable for tasks that require nuanced understanding of text. [12], [13]. Existing work explores document–label grouping or text–label matching [14], [15], refines positive/negative pair construction via label similarity [16]. In parallel, incorporating label embeddings can improve multi-label classification by leveraging label semantics and correlations [17]–[22]. However, for hierarchical multi-label patent classification, two gaps remain: (i) fine-grained correlations across different hierarchical paths and levels are not adequately encoded in similarity measurement and contrastive constraints, and (ii) complementary signals between patent text and label semantics are underexploited due to insufficient fine-grained text–label interaction modeling.

To address these gaps, we propose CLCA, a patent classification framework based on Contrastive Learning and Co-Attention. We design a hierarchy-aware supervised contrastive objective that defines pairwise similarity using both label hierarchy and label similarity, assigning higher weights to sample pairs sharing deeper-level labels to learn hierarchy-sensitive and discriminative patent text representations. We further introduce a co-attention fusion module to explicitly model bidirectional interactions between label semantics and patent text, strengthening fine-grained alignment and improving decision boundary characterization.

Our contributions are summarized as follows:

- We propose a hierarchy-aware supervised contrastive objective for hierarchical multi-label patent classification, where pairwise relatedness is computed from label-path similarity and used to weight contrastive constraints, preventing semantically close samples under shared ancestors from being treated as hard negatives.
- We design a co-attention-based text–label fusion module that explicitly models bidirectional interactions between patent text features and label semantics, strengthening fine-grained alignment for hierarchical multi-label prediction.
- Extensive experiments on two real-world patent datasets in different languages show that CLCA consistently outperforms strong baselines across multiple metrics and Top-K settings.

II. Methodology

CLCA (Fig. 1) comprises four modules: Patent Encoding (PE), Hierarchical Multi-label Contrastive Learning (HMCL), Text–Label Cross (TLC), and Classification Prediction (CP). PE encodes patent texts and IPC labels into embeddings; HMCL refines text representations under hierarchical label constraints via contrastive learning; TLC performs co-attentive text–label fusion to capture complementary signals; and CP produces category-wise prediction scores from the fused representations.

A. Patent Encoding (PE)

Given a patent text t_k, we first map each token to a d-dimensional *word2vec* embedding and obtain $H_k = [h_1^k, \ldots, h_n^k]$. We then employ a Bi-LSTM to encode the token sequence and produce contextualized text features:

$$\mathrm{DSA} = \text{Bi-LSTM}(H), \quad \mathrm{DSA} \in \mathbb{R}^{n \times f}. \tag{1}$$

We model the IPC taxonomy as a hierarchy $\mathcal{T} = \{C, R\}$ [9], where C^l denotes the set of IPC codes at level l, and $R = \{R^{1,2}, \ldots, R^{L-1,L}\}$ encodes parent–child relations across adjacent levels. Specifically, $R^{l,l+1} \in \mathbb{R}^{|C^l| \times |C^{l+1}|}$ is a binary matrix where $R^{l,l+1}_{i,j} = 1$ if c_i^l is the parent of c_j^{l+1}, and 0 otherwise. For a code c_i^q at level q, we define its horizontal neighborhood $S^H_{c_i^q}$ as the set of level-q codes sharing the same parent with c_i^q, and its vertical neighborhood $S^V_{c_i^q}$ as the union of its parent at level $q-1$ and its children at level $q+1$, capturing the immediate hierarchical context. Embeddings of IPC codes within the hierarchical taxonomy as $\{E^1, \ldots, E^L\}$, where $E^l \in \mathbb{R}^{|C^l| \times f}$. To incorporate taxonomy structure, for each label c_i^q we aggregate information from (i) *horizontal* neighbors (same-level siblings sharing the same parent) and (ii) *vertical* neighbors (its parent and children across adjacent levels). We compute label affinities by inner product and normalize them over the corresponding neighbor set:

$$\beta^{q,p}_{i,j} = E^q_i E^{p\,T}_j, \tag{2}$$

$$\widehat{\beta}^{q,p}_{i,j} = \frac{\exp(\beta^{q,p}_{i,j})}{\sum\limits_{c_e^p \in S^*_{c_i^q}} \exp(\beta^{q,p}_{i,e})}, \quad * \in \{V, H\},\ p \in [q-1, q, q+1], \tag{3}$$

where $* \in \{V, H\}$ indicates the vertical or horizontal perspective, respectively. Then obtain aggregated representations:

$$\widehat{E}^{q,*}_i = \sum_{c_j^p \in S^*_{c_i^q}} \widehat{\beta}^{q,p}_{i,j} E^p_j. \tag{4}$$

We fuse the vertical and horizontal views via an MLP:

$$H^{MP}_q = f\big([\widehat{E}^{q,V}, \widehat{E}^{q,H}]\big), \tag{5}$$

and further concatenate ancestor-level representations to obtain the final label semantics for level q:

$$VL = g\big(\|_{i=1}^{q} H^{MP}_i\big), \quad VL \in \mathbb{R}^{|C^q| \times f}, \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ are implemented as two-layer MLPs.

B. Hierarchical Multi-Label Contrastive Learning (HMCL)

As illustrated in Fig. 2, we extend multi-label contrastive learning to better fit hierarchical multi-label patent classification. Prior MLCL-KNN [16] estimates sample relatedness mainly by the number of shared labels, which can be misleading under a hierarchy: two patents may be treated as a negative pair while still sharing the same ancestor, implying strong semantic proximity. Pushing such samples far apart distorts hierarchical semantics. We therefore redefine inter-sample relatedness using the label taxonomy and introduce a hierarchy-aware supervised contrastive loss.

For contrastive learning, we obtain a sample-level representation by mean pooling over tokens, $\bar{\mathrm{DSA}}_i = \mathrm{MeanPool}(\mathrm{DSA}_i) \in \mathbb{R}^f$, and feed it into a projection head:

$$Z_i = W_2\, \mathrm{ReLU}(W_1\, \bar{\mathrm{DSA}}_i). \tag{7}$$

Let the taxonomy have L levels and represent each label as a root-to-leaf path $l = (l^1, \ldots, l^L)$. For two label paths l_a and l_b, we measure their similarity by the normalized longest common prefix:

$$s(l_a, l_b) = \frac{\ell(l_a, l_b)}{L}. \tag{8}$$

For samples i and j with m and n ground-truth labels, respectively, we aggregate label-path similarities and normalize:

$$\sigma_{ij} = \frac{\sum_{a=1}^{m} \sum_{b=1}^{n} s(l_a, l_b)}{m\, n}, \tag{9}$$

where $\sigma_{ij} \in [0, 1]$ serves as a hierarchy-aware weight indicating the semantic relatedness between the two samples.

The weighted NT-Xent loss for a pair (i, j) is defined as:

$$L^{ij}_{con} = -\sigma^{\alpha}_{ij} \log \frac{\exp\big(\mathrm{sim}(Z_i, Z_j)/\tau\big)}{\sum\limits_{k=1, k\neq i}^{b} \exp\big(\mathrm{sim}(Z_i, Z_k)/\tau\big)}, \tag{10}$$

where τ is the temperature and α controls the sharpness of the hierarchy-aware weighting. Finally, we compute the batch contrastive loss by averaging over all ordered pairs:

$$L^{CON} = \frac{1}{b^2} \sum_{i=1}^{b} \sum_{j=1, j\neq i}^{b} L^{ij}_{con}. \tag{11}$$

C. Text–Label Cross (TLC)

TLC explicitly models bidirectional interactions between patent text features DSA and label semantics VL via co-attention, enabling label-informed text encoding and text-informed label refinement. We first compute a normalized text–label relevance matrix $\mathrm{NVC} \in \mathbb{R}^{n \times |C^q|}$:

$$\mathrm{NVC} = \mathrm{softmax}\left(\frac{\mathrm{DSA} \cdot \mathrm{VL}^T}{\|\mathrm{DSA}\| \times \|\mathrm{VL}\|}\right). \tag{12}$$

Using NVC, we derive label-aware text features:

$$\mathrm{LA} = \mathrm{NVC} \cdot \mathrm{VL}. \tag{13}$$

We then construct two cross-attention blocks with residual connections: (i) LA-co, where text attends to label features,

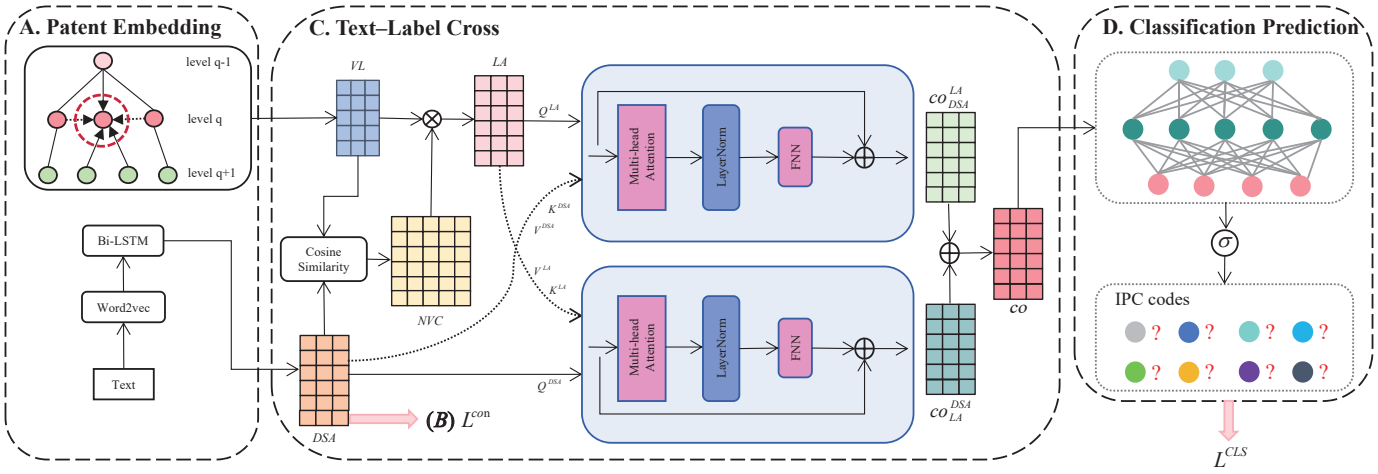

Fig. 1. Overall architecture of CLCA. The framework comprises four distinct modules. Taking a given patent text and IPC taxonomy as inputs, it performs hierarchical multi-label contrastive learning and text–label interaction, followed by the prediction layer, to output the labels assigned to the patent.

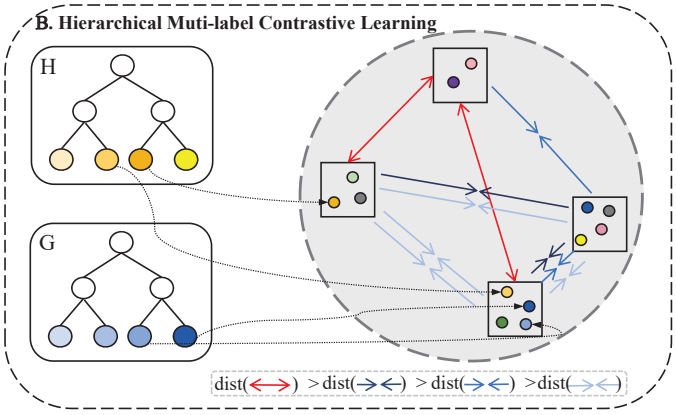

Fig. 2. Hierarchical Multi-Label Contrastive Learning, different colors are used to indicate the strength of contrastive forces: lighter colors correspond to weaker push–pull interactions between two samples.

and (ii) DSA-co, where label-aware features attend back to the text. With linear projections producing Q, K, V for DSA and LA, the co-attentive representations are:

$$\mathrm{co}_{\mathrm{LA}}^{\mathrm{DSA}} = \mathrm{DSA}{+}\mathrm{FNN}\Big(\mathrm{LN}\big(\mathrm{MultiHead}(Q^{\mathrm{DSA}}, K^{\mathrm{LA}}, V^{\mathrm{LA}})\big)\Big), \tag{14}$$

$$\mathrm{co}_{\mathrm{DSA}}^{\mathrm{LA}} = \mathrm{LA}{+}\mathrm{FNN}\Big(\mathrm{LN}\big(\mathrm{MultiHead}(Q^{\mathrm{LA}}, K^{\mathrm{DSA}}, V^{\mathrm{DSA}})\big)\Big), \tag{15}$$

where $\mathrm{MultiHead}(\cdot)$ is multi-head attention, $\mathrm{LN}(\cdot)$ is layer normalization, and $\mathrm{FNN}(\cdot)$ is a position-wise feed-forward network. Finally, we fuse the two views with a convex combination:

$$\mathrm{CO} = \mu \cdot \mathrm{co}_{\mathrm{LA}}^{\mathrm{DSA}} + (1-\mu) \cdot \mathrm{co}_{\mathrm{DSA}}^{\mathrm{LA}}. \tag{16}$$

D. Classification Prediction (CP)

Given the fused features CO from TLC, we predict IPC codes at level q using a two-layer MLP decoder followed by a sigmoid:

$$\hat{\mathbf{Y}}_i = \sigma(g(\mathrm{CO}_i)), \quad \hat{\mathbf{Y}}_i \in \mathbb{R}^{|C^q|}. \tag{17}$$

We treat patent classification as multi-label learning with binary relevance and optimize the binary cross-entropy loss:

$$L^{CLS} = -\sum_{i=1}^{N}\sum_{j=1}^{|C^q|}\Big[Y_{i,j}\log(\hat{Y}_{i,j}) + (1-Y_{i,j})\log(1-\hat{Y}_{i,j})\Big], \tag{18}$$

where $Y_{i,j} \in \{0,1\}$ indicates whether label j is assigned to patent t_i.

Finally, we jointly train the classifier with the hierarchical contrastive objective:

$$L^{total} = L^{CLS} + \lambda L^{CON}, \tag{19}$$

where λ balances the classification and contrastive losses.

III. Experiments

In this section, we conduct extensive experiments on real-world patent text datasets to evaluate the performance of our model.

TABLE I
Statistics of the Two Datasets

Datasets	#C	#patents			#IPC codes		
		#train	#validate	#test	L1	L2	L3
USPTO-200K	513	152,349	24,335	23,675	8	125	622
CNPTD-200K	1,443	120,130	40,128	39,807	8	124	608

A. Dataset Description

We evaluate CLCA on two benchmark patent datasets released by Zou et al. [9]. Table. I reports the dataset statistics, including the numbers of documents in train/validation/test splits, the total number of IPC codes, and the label cardinalities at each hierarchy level (L1–L3).

USPTO-200K contains about 200K patents from the USPTO. Patents published in 2010–2017, 2018, and 2019 are used for training, validation, and testing, respectively.

CNPTD-200K consists of around 200K patents from Chinese technology companies. We adopt the same time-based split as USPTO-200K: 2010–2017 for training, 2018 for validation, and 2019 for testing.

B. Compared Methods

We compare CLCA with nine representative baselines covering text-only modeling, hierarchical multi-label learning, and multi-source patent classification. Text-only methods include Word2Vec [23], Bi-LSTM [24], BERT [6], DeepPatent [3], and Patentsberta [8]. Hierarchical multi-label frameworks include HARNN [17] and HmcNet [19]. DyGFormer [25] is a Transformer-based model that captures temporal interactions in assignee–patent graphs. PatCLS [9] is a recent state-of-the-art patent classifier that integrates text embeddings, adaptive IPC correlation learning, and assignees' historical behaviors.

C. Evaluation Metrics

We evaluate all models using Precision@K, Recall@K, F1@K, and NDCG@K. For each patent t_i, let $\hat{S}_i$ be the Top-K predicted label set and S_i the ground-truth label set. All metrics are averaged over the test set.

$$\text{Precision@K}(t_i) = \frac{|\hat{S}_i \cap S_i|}{|\hat{S}_i|}, \quad \text{Recall@K}(t_i) = \frac{|\hat{S}_i \cap S_i|}{|S_i|}. \tag{20}$$

$$\text{F1@K}(t_i) = \frac{2\,\text{Precision@K}(t_i)\,\text{Recall@K}(t_i)}{\text{Precision@K}(t_i) + \text{Recall@K}(t_i) + \epsilon}. \tag{21}$$

where ϵ is a small constant for numerical stability.

Let $\hat{S}_i^k$ denote the k-th ranked predicted label and $\delta(\hat{S}_i^k, S_i)$ be 1 if $\hat{S}_i^k \in S_i$ and 0 otherwise. We compute:

$$\text{NDCG@K}(t_i) = \frac{\sum_{k=1}^{K} \dfrac{\delta(\hat{S}_i^k, S_i)}{\log_2(k+1)}}{\sum_{k=1}^{\min(K,|S_i|)} \dfrac{1}{\log_2(k+1)}}. \tag{22}$$

D. Experimental Settings

We use third-level IPC codes (L3) as ground-truth labels. For each patent, we truncate the input to the first 100 words in the "Summary" section. Models are trained with Adam,learning rate $= 2 \times 10^{-4}$ and dropout= 0.5. For contrastive learning, we set the temperature $\tau = 0.5$, loss weight $\lambda = 0.1$ and $\alpha = 3$. We train for up to 200 epochs with early stopping (patience=10) based on validation performance, and report test results using the best checkpoint. All experiments are run on an NVIDIA RTX 3090 (Ti) GPU. For all models, we conduct five independent runs under different random seeds and report the average test performance.

E. Performance Comparison

Table. II and Table. III report Top-K results ($K = 1, 3, 5$) on USPTO-200K and CNPTD-200K.

USPTO-200K: CLCA achieves the best performance across all metrics and all K. Compared with PatCLS, it consistently improves F1 and NDCG, with the most notable gains on Recall and NDCG. For example, at K=3, CLCA improves Recall@3 and NDCG@3 by 2.70% and 2.63% (relative), respectively; at K=5, it further yields relative gains of 1.88% in Recall@5 and 2.31% in NDCG@5. These results indicate better label coverage and ranking quality within the top-K predictions, rather than a trivial precision–recall trade-off. The larger margin over text-only baselines suggests that modeling hierarchical constraints and text–label semantics is critical for high-quality top-K prediction.

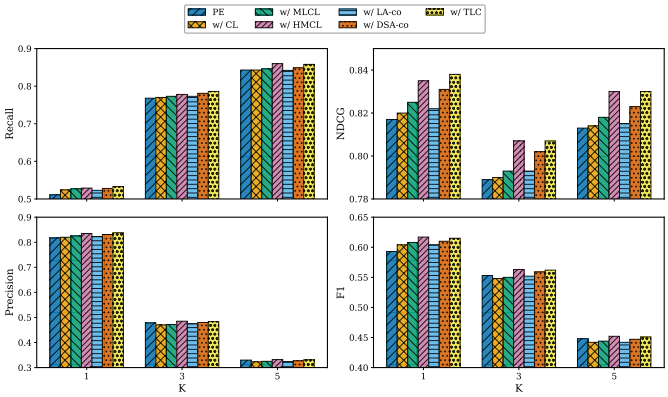

Fig. 3. Different module combinations on the USPTO-200K dataset.

CNPTD-200K: Although CNPTD-200K is more challenging, CLCA still outperforms all baselines under all K. The gains over PatCLS are consistent and mainly reflected in Recall and NDCG. Specifically, CLCA achieves 1.62% and 1.60% relative improvements in Recall@3 and NDCG@3, and 1.15% and 1.28% relative improvements in Recall@5 and NDCG@5, respectively, demonstrating robust effectiveness across datasets with different data distributions.

F. Ablation Study

1) Impact of Different Components on Multi-Label Patent Classification: To assess the contributions of hierarchy-aware contrastive learning (HMCL) and text–label fusion (TLC), we evaluate four variants: PE (base model), PE+HMCL, PE+TLC, and the full model CLCA. Table. IV reports Precision, Recall, F1, and NDCG at $K \in \{1, 3, 5\}$ on USPTO-200K. Overall, CLCA achieves the best results under all K. Compared with PE+HMCL, CLCA consistently improves all metrics, with absolute gains ranging from 0.5 to 2.0 points (averaging ≈1.0 point), and the most notable improvement on Recall@3 (+2.0) and NDCG@3 (+1.3), highlighting the benefit of fusing label semantics with text via co-attention. Compared with PE+TLC, CLCA also yields consistent gains of 0.5–1.3 points (averaging ≈1.0 point) across metrics, suggesting that hierarchy-aware contrastive learning provides complementary supervision for learning more discriminative text representations. Together, these results indicate that HMCL and TLC are both effective individually and complementary when combined.

2) Effect of Hierarchical Multi-Label Contrastive Learning: We further isolate the impact of HMCL by fixing the backbone

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON USPTO-200K DATASET.

Methods	$K=1$				$K=3$				$K=5$			
	P	R	F1	NDCG	P	R	F1	NDCG	P	R	F1	NDCG
Word2Vec	0.763	0.462	0.542	0.763	0.456	0.713	0.521	0.735	0.318	0.792	0.428	0.760
Bi-LSTM	0.799	0.485	0.568	0.799	0.471	0.736	0.538	0.763	0.327	0.813	0.439	0.786
Bert	0.792	0.480	0.562	0.792	0.468	0.729	0.533	0.756	0.325	0.805	0.436	0.779
DeepPatent	0.800	0.484	0.567	0.800	0.471	0.735	0.538	0.762	0.327	0.812	0.439	0.785
HARNN	0.809	0.489	0.573	0.808	0.480	0.746	0.547	0.773	0.334	0.826	0.448	0.798
HmcNet	0.818	0.512	0.593	0.818	0.476	0.766	0.550	0.787	0.327	0.839	0.445	0.810
DyGFormer	0.831	0.520	0.603	0.831	0.482	0.776	0.558	0.798	0.332	0.849	0.450	0.821
Patentsberta	0.805	0.503	0.584	0.805	0.470	0.760	0.545	0.777	0.326	0.837	0.443	0.803
PatCLS	0.833	0.521	0.605	0.833	0.484	0.777	0.559	0.799	0.332	0.852	0.451	0.823
Ours	**0.849***	**0.538***	**0.622***	**0.849***	**0.493***	**0.798***	**0.572***	**0.820***	**0.337***	**0.868***	**0.459***	**0.842***

* indicates statistically significant improvement over the best baseline (paired t-test, $p < 0.05$).

TABLE III
PERFORMANCE OF DIFFERENT MODELS ON CNPTD-200K DATASET.

Methods	$K=1$				$K=3$				$K=5$			
	P	R	F1	NDCG	P	R	F1	NDCG	P	R	F1	NDCG
Word2Vec	0.585	0.483	0.514	0.585	0.304	0.699	0.408	0.648	0.205	0.772	0.313	0.679
Bi-LSTM	0.641	0.534	0.566	0.641	0.324	0.752	0.437	0.702	0.216	0.818	0.331	0.731
Bert	0.557	0.458	0.488	0.557	0.300	0.694	0.404	0.633	0.207	0.780	0.316	0.671
DeepPatent	0.658	0.547	0.566	0.659	0.335	0.722	0.451	0.722	0.222	0.839	0.341	0.751
HARNN	0.625	0.520	0.552	0.625	0.315	0.731	0.425	0.683	0.210	0.799	0.323	0.713
HmcNet	0.679	0.574	0.606	0.679	0.339	0.800	0.460	0.746	0.224	0.863	0.345	0.774
DyGFormer	0.682	0.577	0.609	0.682	0.337	0.798	0.459	0.746	0.226	0.861	0.344	0.774
PatCLS	0.687	0.582	0.614	0.687	0.340	0.804	0.462	0.752	0.224	0.868	0.346	0.780
Ours	**0.697***	**0.591***	**0.623***	**0.697***	**0.347***	**0.817***	**0.471***	**0.764***	**0.228***	**0.878***	**0.351***	**0.790***

Note: Patentsberta is pretrained on English corpora and is therefore ill-suited for Chinese patent texts in this task.

TABLE IV
ABLATION STUDY OF CLCA ON USPTO-200K.

Methods	$K=1$				$K=3$				$K=5$			
	P	R	F1	NDCG	P	R	F1	NDCG	P	R	F1	NDCG
PE	0.817	0.511	0.593	0.817	0.478	0.768	0.553	0.789	0.330	0.843	0.448	0.813
PE+HMCL	0.835	0.529	0.617	0.835	0.485	0.778	0.563	0.807	0.332	0.860	0.452	0.830
PE+TLC	0.838	0.533	0.615	0.838	0.483	0.786	0.562	0.807	0.331	0.858	0.451	0.830
Ours	**0.849**	**0.538**	**0.622**	**0.849**	**0.493**	**0.798**	**0.572**	**0.820**	**0.337**	**0.868**	**0.459**	**0.842**

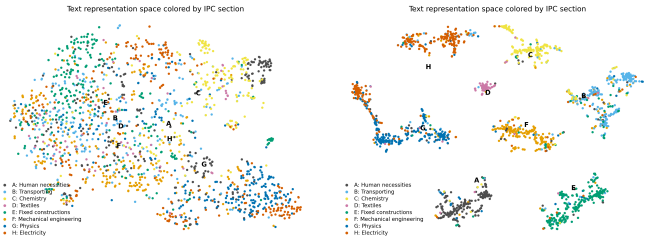

Fig. 4. t-SNE visualization of patent text embeddings, where the left plot corresponds to PE and the right plot corresponds to PE+HMCL. Each color denotes embeddings of patent texts associated with a different label.

to PE and comparing three contrastive objectives: w/ CL (positives share identical label sets), w/ MLCL (positives weighted by the number of shared labels), and w/ HMCL (our hierarchy-aware weighting based on label-path similarity; Eq. (8)–(11)). Fig. 3 shows that w/ HMCL consistently improves Recall and NDCG at $K = \{1, 3, 5\}$ over both variants, indicating that explicitly accounting for hierarchical relatedness yields better label coverage and ranking quality. The t-SNE visualization in Fig. 4 further supports this observation: embeddings learned without HMCL are highly entangled across IPC sections, whereas HMCL produces clearer and more separated clusters, implying improved inter-class separability.

3) Effect of Text–Label Cross: To validate the proposed text–label interaction, we compare: PE (simple concatenation of text and label features), w/ LA-co (adding label-to-text co-attention), w/ DSA-co (adding text-to-label co-attention), and the full w/ TLC. As shown in Fig. 3, PE performs worst, suggesting that coarse feature stacking cannot effectively exploit complementary signals. Introducing either unidirectional branch improves Recall and NDCG, while the bidirectional

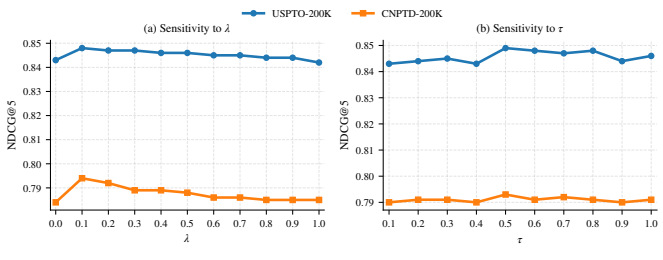

Fig. 5. Effects of λ (left) and τ (right) on NDCG@5 on the USPTO-200K and CNPTD-200K dataset.

co-attention in w/ TLC achieves the best performance across K, demonstrating that mutual alignment between text cues and label semantics is crucial for high-quality top-K ranking.

G. Parameter Sensitivity Analysis

As Eq. (19) and Eq. (10) show, λ and τ are the contrastive-loss weight and temperature, respectively. Fig. 5 reports NDCG@5 under different λ and τ. Performance peaks at a small λ and degrades as λ increases, suggesting that overly strong contrastive regularization may separate semantically related samples with slight label differences. For τ, NDCG@5 exhibits a unimodal trend and reaches the best value at $\tau = 0.5$. A relatively small temperature emphasizes informative pairs and alleviates the scarcity of highly similar pairs in patent data, leading to better ranking quality.

IV. Conclusion

Patent classification is challenging due to highly technical language with coined terminology and the need to align patent text with hierarchical IPC labels. To address these challenges, we proposed CLCA, which learns hierarchy-sensitive text representations via a supervised contrastive objective weighted by hierarchical label-path similarity and further strengthens fine-grained text–label alignment through a co-attention fusion module. Experiments on English and Chinese patent datasets demonstrate that CLCA consistently outperforms strong baselines across multiple metrics and Top-K settings, validating the effectiveness of jointly leveraging hierarchical label structure and label semantics for hierarchical multi-label patent classification.

V. Acknowledgments

This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant Nos. 2026C02A1242 and 2026C01015.

References

[1] L. Fang, L. Zhang, H. Wu, T. Xu, D. Zhou, and E. Chen, "Patent2vec: Multi-view representation learning on patent-graphs for patent classification," *World Wide Web*, vol. 24, no. 5, pp. 1791–1812, 2021.

[2] L. Jiang and S. M. Goetz, "Natural language processing in the patent domain: a survey," *Artificial Intelligence Review*, vol. 58, no. 7, p. 214, 2025.

[3] S. Li, J. Hu, Y. Cui, and J. Hu, "Deeppatent: patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, 2018.

[4] M. Sofean, "Deep learning based pipeline with multichannel inputs for patent classification," *World Patent Information*, vol. 66, p. 102060, 2021.

[5] A. H. Roudsari, J. Afshar, S. Lee, and W. Lee, "Comparison and analysis of embedding methods for patent documents," in *2021 IEEE international conference on big data and smart computing (BigComp)*. IEEE, 2021, pp. 152–155.

[6] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning bert language model," *World Patent Information*, vol. 61, p. 101965, 2020.

[7] A. Haghighian Roudsari, J. Afshar, W. Lee, and S. Lee, "Patentnet: multi-label classification of patent documents using deep learning based language understanding," *Scientometrics*, vol. 127, no. 1, pp. 207–231, 2022.

[8] H. Bekamiri, D. S. Hain, and R. Jurowetzki, "Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert," *Technological Forecasting and Social Change*, vol. 206, p. 123536, 2024.

[9] T. Zou, L. Yu, J. Ye, L. Sun, B. Du, and D. Wang, "Adaptive taxonomy learning and historical patterns modeling for patent classification," *ACM Transactions on Information Systems*, vol. 42, no. 6, pp. 1–24, 2024.

[10] L. Zhang, W. Liu, Y. Chen, and X. Yue, "Reliable multi-view deep patent classification," *Mathematics*, vol. 10, no. 23, p. 4545, 2022.

[11] E. Kamateri, M. Salampasis, and K. Diamantaras, "An ensemble framework for patent classification," *World Patent Information*, vol. 75, p. 102233, 2023.

[12] Y. Liang, T. Tohti, and A. Hamdulla, "Contrastive classification: A label-independent generalization model for text classification," *Expert Systems with Applications*, vol. 245, p. 123130, 2024.

[13] A. Audibert, A. Gauffre, and M.-R. Amini, "Multi-label contrastive learning: A comprehensive study," *arXiv preprint arXiv:2412.00101*, 2024.

[14] J. Xiong, Y. Li, and X. Niu, "Big: A bidirectional group-wise contrastive learning method for multi-label text classification," *Expert Systems with Applications*, p. 127977, 2025.

[15] H. Ye, R. Sunderraman, and S. Ji, "Matchxml: an efficient text-label matching framework for extreme multi-label text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 9, pp. 4781–4793, 2024.

[16] J. Zhang, Y. Li, F. Shen, Y. He, H. Tan, and Y. He, "Hierarchical text classification with multi-label contrastive learning and knn," *Neurocomputing*, vol. 577, p. 127323, 2024.

[17] W. Huang, E. Chen, Q. Liu, Y. Chen, Z. Huang, Y. Liu, Z. Zhao, D. Zhang, and S. Wang, "Hierarchical multi-label text classification: An attention-based recurrent network approach," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1051–1060.

[18] S. Ran, X. Li, B. Zhao, Y. Jiang, X. Yang, and C. Cheng, "Label correlation embedding guided network for multi-label ecg arrhythmia diagnosis," *Knowledge-Based Systems*, vol. 270, p. 110545, 2023.

[19] W. Huang, E. Chen, Q. Liu, H. Xiong, Z. Huang, S. Tong, and D. Zhang, "Hmcnet: A general approach for hierarchical multi-label classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8713–8728, 2022.

[20] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, p. 101718, 2022.

[21] G. Sun, Y. Cheng, F. Dong, L. Wang, D. Zhao, Z. Zhang, and X. Tong, "Multi-label text classification model integrating label attention and historical attention," *Knowledge-Based Systems*, vol. 296, p. 111878, 2024.

[22] Q. Wang, J. Zhu, H. Shu, K. O. Asamoah, J. Shi, and C. Zhou, "Gudn: A novel guide network with label reinforcement strategy for extreme multi-label text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 4, pp. 161–171, 2023.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[24] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[25] L. Yu, L. Sun, B. Du, and W. Lv, "Towards better dynamic graph learning: New architecture and unified library," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 686–67 700, 2023.